Solution for the properties of a clustered network

Juyong Park and M. E. J. Newman
Department of Physics, University of Michigan, Ann Arbor, MI 48109–1120, U.S.A.

We study Strauss's model of a network with clustering and present an analytic mean-field solution which is exact in the limit of large network size. Previous computer simulations have revealed a degenerate region in the model's parameter space in which triangles of adjacent edges clump together to form unrealistically dense subgraphs, and perturbation calculations have been found to break down in this region at all orders. Our solution shows that this region corresponds to a classic symmetry-broken phase and that the onset of the degeneracy corresponds to a first-order phase transition in the density of the network.

I. INTRODUCTION

The last few years have seen a surge of interest within the scientific community in the properties of networks of various kinds [1, 2, 3]. In parallel with empirical studies of real-world networks such as the Internet [4], the worldwide web [5, 6], biological networks [7, 8], and social networks [9], researchers have developed theoretical models and mathematical tools to explain the rich structure and nontrivial characteristics that large-scale networks exhibit.

The most fundamental of network models may be the Bernoulli random graph [10] (also sometimes called the Erdős–Rényi model after two well-known mathematicians who were among the first to study it [11]). In this model, n identical vertices are joined together in pairs by edges, each possible edge appearing with independent probability p for a total of $\binom{n}{2}p$ edges on average. This model can be thought of as a special case of the much larger class of *exponential random graphs*, which is the class of ensembles of graphs that maximize ensemble entropy under a given set of constraints (usually imposed by observations of the properties of an actual network in the real world) [12]. The appropriate constraint for the Bernoulli random graph is a constraint on the total number of edges in the graph.

The exponential random graph model defines a probability distribution over a specified set of possible graphs such that the probability $P(G)$ of a particular graph G is proportional to $\mathrm{e}^{-H(G)}$, where

$$H(G) = \sum_i \theta_i m_i(G). \tag{1}$$

$H(G)$ is called the *graph Hamiltionian*, $\{m_i\}$ is the set of observables upon which the relevant constraints act, and $\{\theta_i\}$ is a set of real-valued conjugate fields which we can vary so as to match the properties of the model to the real-world network under consideration. Exact or approximate solutions of average properties of the ensemble are possible for a variety of graph Hamiltonians, including graphs with arbitrary degree distributions, directed-graph models with reciprocity [12], the so-called 2-star model [13], and others [14].

In this paper we give a solution of a particular famous exponential random graph model, the clustering model of Strauss [15]. This model mimics the phenomenon of network transitivity or clustering, which has been much discussed in the networks literature [9, 16, 17, 18]. The model was originally proposed in 1981 and has recently attracted the attention of the physics community [19, 20], where the question of how properly to model transitivity has proved a persistent stumbling block for theorists.

II. STRAUSS'S MODEL OF CLUSTERING

Strauss's model is simple to define. The appropriate graph observables are the number of edges $m(G)$ and the number of triangles $t(G)$, so that the Hamiltonian can be written

$$\begin{aligned} H(G) &= \theta m(G) - \alpha t(G) \\ &= \theta \sum_{i<j} \sigma_{ij} - \alpha \sum_{i<j<k} \sigma_{ij}\sigma_{jk}\sigma_{ki}, \end{aligned} \tag{2}$$

where $\sigma_{ij} = \sigma_{ji}$ is an element of the *adjacency matrix* having value 1 if an edge exists between vertices i and j and 0 otherwise. When $\alpha > 0$, this Hamiltonian encourages the formation of triangles in the network by assigning lower "energy" to graphs with many triangles.

Although the Hamiltonian seems simple enough, Strauss found via numerical simulations that the model sometimes behaved strangely, developing in certain parameter regimes a "degenerate state," a condensed phase in which many triangles form but tend to stick together in local regions of the graph, rather than spreading uniformly over it. Recently Burda *et al.* [19] have performed a perturbation theoretic analysis of the model [23], finding that the formation of this condensed phase corresponds to a point at which the perturbation series breaks down at all orders simultaneously. The nature of this point and of the condensed phase however has not been well understood and a complete solution of the model has been lacking. In the next section, we present a solution of the model based on a mean-field approach which we believe to be exact for all parameter values in the limit of large system size. Using this solution, we show that the model possess a classic second-order phase transition between a high-symmetry regime and a symmetry-broken one, with a line of first-order transitions between

states of high and low density in the symmetry-broken regime. The formation of the "condensed phase" observed by Strauss corresponds precisely to the first-order transition from low to high density.

III. ANALYSIS

A. Mean-field solution

Let H_{ij} be the sum of all terms in the Hamiltonian, Eq. (2), that involve σ_{ij}:

$$H_{ij} = \theta\sigma_{ij} - \alpha \sum_{k\neq i,j} \sigma_{ij}\sigma_{jk}\sigma_{ki} = \sigma_{ij}\big(\theta - \alpha \sum_{k\neq i,j} \sigma_{jk}\sigma_{ki}\big), \tag{3}$$

and let H' be the remaining terms, so that $H = H_{ij} + H'$. The mean value $\langle\sigma_{ij}\rangle$ of σ_{ij} can then be written as

$$\begin{aligned}\langle\sigma_{ij}\rangle &= 0\times P(\sigma_{ij}=0) + 1\times P(\sigma_{ij}=1) \\ &= \frac{1}{Z}\sum_{\{\sigma\}} \mathrm{e}^{-H} \frac{\mathrm{e}^{-H_{ij}(\sigma_{ij}=1)}}{\mathrm{e}^{-H_{ij}(\sigma_{ij}=0)} + \mathrm{e}^{-H_{ij}(\sigma_{ij}=1)}} \\ &= \left\langle \frac{1}{\mathrm{e}^{\theta-\alpha\sum_{k\neq i,j}\sigma_{jk}\sigma_{ki}}+1} \right\rangle. \end{aligned} \tag{4}$$

where $Z = \sum_G \mathrm{e}^{-H(G)}$ is the partition function. Here $\langle\ldots\rangle$ indicates the average within the ensemble, and the derivation so far has been exact.

By analogy with spin models, let us call the expression within the brackets in Eq. (3) the *local field* coupled to spin σ_{ij}. The mean-field approximation involves replacing the spin variables in the local field with their ensemble averages, which in this case means $\sigma_{jk}\sigma_{ki} \to q \equiv \langle\sigma_{jk}\sigma_{ki}\rangle$. Defining also the *connectance* $p \equiv \langle\sigma_{ij}\rangle$, we now have

$$p = \frac{1}{\mathrm{e}^{\theta-\alpha(n-2)q}+1} = \tfrac{1}{2}\big[1-\tanh\big(\tfrac{1}{2}\theta - \tfrac{1}{2}\alpha(n-2)q\big)\big]. \tag{5}$$

Now we set up an equation for q via a similar method. Noting that $\sigma_{ik}\sigma_{kj} = 1$ only when both $\sigma_{ik} = 1$ and $\sigma_{kj} = 1$, we can write:

$$\begin{aligned} q &\equiv \langle\sigma_{ik}\sigma_{kj}\rangle = \\ &\left\langle \frac{\mathrm{e}^{\alpha\sigma_{ij}}}{\big(\mathrm{e}^{\theta-\alpha\sum_l \sigma_{il}\sigma_{lk}}+1\big)\big(\mathrm{e}^{\theta-\alpha\sum_l \sigma_{kl}\sigma_{lj}}+1\big) + (\mathrm{e}^{\alpha\sigma_{ij}}-1)} \right\rangle \\ &= \frac{1+(\mathrm{e}^{\alpha}-1)p}{\big(\mathrm{e}^{\theta-\alpha(n-3)q}+1\big)^2 + (\mathrm{e}^{\alpha}-1)p}, \end{aligned} \tag{6}$$

where in the final line we have made the mean-field approximation again, and made use of the property that $\mathrm{e}^{\alpha\sigma_{ij}} = 1 + (\mathrm{e}^{\alpha}-1)\sigma_{ij}$, since $\sigma_{ij} = 0$ or 1.

We now have two equations in two unknowns which can be solved by substituting (5) into (6) to give a self-consistency condition on q:

$$\begin{aligned} q &= \frac{\mathrm{e}^{\theta-\alpha(n-2)q} + \mathrm{e}^{\alpha}}{(\mathrm{e}^{\theta-\alpha(n-3)q}+1)^2(\mathrm{e}^{\theta-\alpha(n-2)q}+1) + (\mathrm{e}^{\alpha}-1)} \\ &\equiv Q(q). \end{aligned} \tag{7}$$

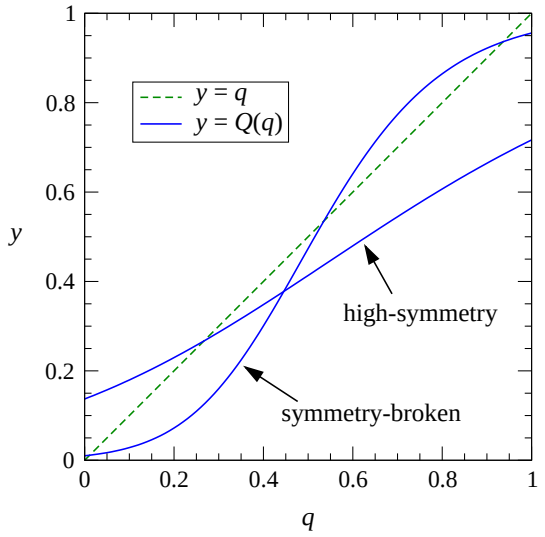

FIG. 1: (Color online) Graphical solutions of $q = Q(q)$. Depending on the values of the parameters θ and α, the line $y = q$ (dashed) intersects with $y = Q(q)$ (solid) either three times or only once. The parameters $(\theta, n\alpha)$ for the two curves shown are $(2.3, 6.0)$ and $(0.5, 2.0)$.

In Fig. 1 we show a plot of the forms $y = q$ and $y = Q(q)$ as functions of q. The intersections of the two curves give the solutions of Eq. (7). As we can see, depending on the values of θ and α, the curves can intersect at either one or three points in the allowed domain $0 < q < 1$. The regime in which there are three solutions corresponds to a symmetric-broken phase with only the outer two solutions being stable (corresponding to minima of the free energy). Thus, the system displays the classic phenomenology of a second-order phase transition, with a critical point separating a high-symmetry phase from a symmetry-broken one having regimes of high- and low-density and an intermediate region of coexistence of the two. In Fig. 2 we show the phase diagram of the system.

Finally we introduce another mean-field equation for $r \equiv \langle\sigma_{ij}\sigma_{jk}\sigma_{ki}\rangle$ which gives the number of triangles in the network:

$$r \equiv \langle\sigma_{ij}\sigma_{jk}\sigma_{ki}\rangle = \frac{\mathrm{e}^{\alpha}}{\big(\mathrm{e}^{\theta-\alpha(n-2)q}+1\big)^3 + (\mathrm{e}^{\alpha}-1)}. \tag{8}$$

In Fig. 3 we compare our solutions for p, q and r with simulation results for a system of size $n = 500$ and, as we can see, the agreement between theory and simulation is excellent.

B. The approximation

As mentioned in the introduction, we believe that the mean-field solution found in the previous section is exact, because in the limit of large system size the system

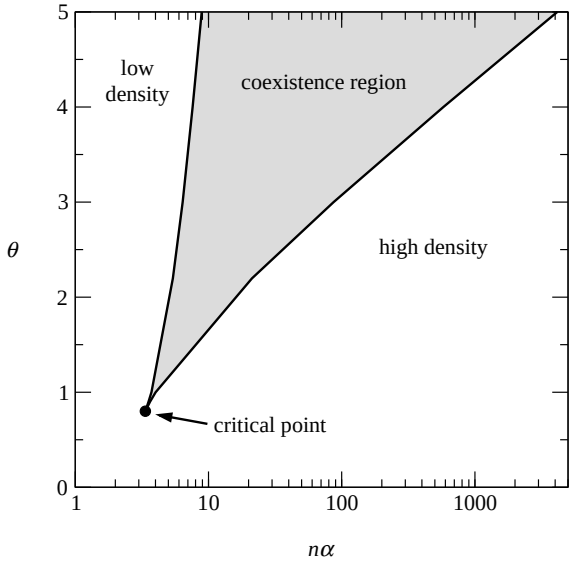

FIG. 2: The phase diagram in the $(n\alpha, \theta)$ space. The shaded area corresponds to the coexistence region in which the system can be in either of two stable states, one of high density and one of low.

becomes fundamentally infinite-dimensional, and mean-field theory is usually exact in the large dimension limit.

In fact, Eq. (4) really makes two approximations. One is the mean-field approximation $\sigma_{jk}\sigma_{ki} \to \langle\sigma_{jk}\sigma_{ki}\rangle$, but we have also assumed that the average of the tanh can be approximated by the tanh of the average. While the first approximation can be justified on the basis of the high effective dimension of the system, the second needs more attention. It can be justified by performing a series expansion of the tanh, applying the mean-field approximation to the series term by term, and then resumming the result again [21, 22]. However, while this method works, it is not as simple as our brief description makes it sound, because the series involves averages over arbitrarily high moments of the graph operators and proving that these terms are negligible requires some care.

Let us rewrite Eq. (4) thus:

$$\begin{aligned} p \equiv \langle\sigma_{ij}\rangle &= \tfrac{1}{2}\bigl(1 + \bigl\langle\tanh(-\tfrac{1}{2}\theta + \tfrac{1}{2}\alpha \sum_{k\neq i,j} S_k)\bigr\rangle\bigr) \\ &= \tfrac{1}{2}\bigl(1 + \bigl\langle\tanh\bigl(\tilde\theta + \tilde\alpha S\bigr)\bigr\rangle\bigr), \end{aligned} \tag{9}$$

where for convenience we have defined $\tilde\theta = -\frac{1}{2}\theta$, $\tilde\alpha = \frac{1}{2}\alpha$, $S_k = \sigma_{ik}\sigma_{kj}$, and $S = \sum_{k\neq i,j} S_k$. Expanding the tanh about $\tilde\theta$, we get

$$\bigl\langle\tanh(\tilde\theta + \tilde\alpha S)\bigr\rangle = \sum_{m=0}^{\infty} \frac{\tanh^{(m)}(\tilde\theta)\tilde\alpha^m}{m!}\langle S^m\rangle. \tag{10}$$

Now, keeping in mind that $S_k^m = \sigma_{ik}^m\sigma_{kj}^m = \sigma_{ik}\sigma_{kj} = S_k$ for any m, since $\sigma_{ij} = 0, 1$, we can write the correlation

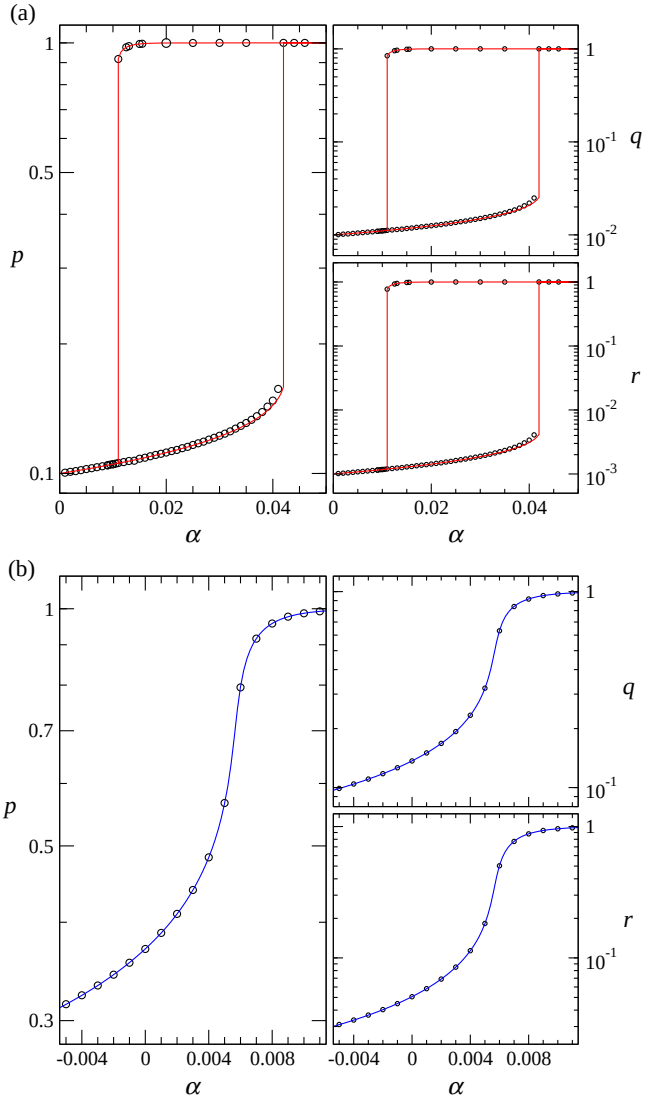

FIG. 3: (Color online) Comparison of our analytic solution (solid lines) and Monte Carlo simulation results (circles) for $p = \langle\sigma_{ij}\rangle$, $q = \langle\sigma_{jk}\sigma_{ki}\rangle$, and $r = \langle\sigma_{ij}\sigma_{jk}\sigma_{ki}\rangle$, for a system of $n = 500$ vertices. The parameter values were (a) $\theta = 2.2$ and (b) $\theta = 0.53$. (See Fig. 2.)

functions $\langle S^m\rangle$ in the form

$$\langle S^m\rangle = \bigl\langle\bigl(\textstyle\sum_k S_k\bigr)\bigl(\sum_k S_k\bigr)\ldots\bigl(\sum_k S_k\bigr)\bigr\rangle \tag{11a}$$

$$= a_{m,1}\bigl\langle S_1 + S_2 + \ldots\bigr\rangle + a_{m,2}\bigl\langle S_1S_2 + S_1S_3 + \ldots\bigr\rangle + a_{m,3}\bigl\langle S_1S_2S_3 + \ldots\bigr\rangle + \ldots \tag{11b}$$

$$= a_{m,1}(n-2)q + a_{m,2}\tbinom{n-2}{2}q^2 + a_{m,3}\tbinom{n-2}{3}q^3 + \ldots \tag{11c}$$

In the last line we have made the assumption that a product $\langle S_1S_2\rangle$ can be approximated as $\langle S_1\rangle\langle S_2\rangle = q^2$, and similarly for higher products. This approximation is of the nature of a mean-field approximation, ignoring correlations between single pairs of spin variables, which

will be of order $1/n$ in the large system size limit where each variable interacts with an arbitrary number of others.

The coefficient $a_{m,l}$ in Eq. (11) is the number of ways of selecting one S_k from each of the m sums in (11a) so that there are l unique indices in the resulting product. It is simple to show that $\sum_{i=1}^{l} \binom{l}{i} a_{m,i} = l^m$ and thus by induction to prove that the exponential generating function for $a_{m,l}$ satisfies

$$g_l(z) = \sum_{m=1}^{\infty} \frac{z^m a_{m,l}}{m!} = (\mathrm{e}^z - 1)^l. \tag{12}$$

Then, by repeated differentiation

$$a_{m,l} = \left[\frac{\partial^m}{\partial z^m} (\mathrm{e}^z - 1)^l \right]_{z=0}. \tag{13}$$

Combining this result with Eq. (11) and taking the limit of large n, we find

$$\begin{aligned} \langle S^m \rangle &= \sum_{l=0}^{\infty} a_{m,l} \tbinom{n-2}{l} q^l \simeq \sum_{l=0}^{\infty} a_{m,l} (n-2)^l q^l / l! \\ &= \left[\frac{\partial^m}{\partial z^m} \sum_{l=0}^{\infty} \frac{(\mathrm{e}^z - 1)^l (n-2)^l q^l}{l!} \right]_{z=0} \\ &= \left[\frac{\partial^m}{\partial z^m} \mathrm{e}^{(n-2)q(\mathrm{e}^z - 1)} \right]_{z=0}. \end{aligned} \tag{14}$$

The differentiation in the last line can be carried out explicitly for any given value of m, but there is no simple closed-form expression for the general case. However, none is needed in the large n limit. Each successive differentiation with respect to z generates an extra factor of $(n-2)q$. But the graphs we are interested in are *dense*, meaning that p, q, and r all tend to finite, non-zero limiting values as $n \to \infty$. Thus $(n-2)q$ is a large quantity and to leading order we need only retain the highest-order term in the derivative, which is simply $\bigl[(n-2)q\bigr]^m$. Thus Eq. (10) becomes

$$\begin{aligned} \bigl\langle \tanh(\tilde\theta + \tilde\alpha S) \bigr\rangle &= \sum_{m=0}^{\infty} \frac{\tanh^{(m)}(\tilde\theta) \tilde\alpha^m}{m!} \bigl\langle S^m \bigr\rangle \\ &= \sum_{m=0}^{\infty} \frac{\tanh^{(m)}(\tilde\theta) \tilde\alpha^m}{m!} \bigl((n-2)q\bigr)^m \\ &= \tanh\bigl(\tilde\theta + \tilde\alpha(n-2)q\bigr) \end{aligned} \tag{15}$$

and

$$p = \tfrac{1}{2}\bigl[1 + \tanh\bigl(\tilde\theta + \tilde\alpha(n-2)q\bigr)\bigr], \tag{16}$$

which is identical with Eq. (5). A similar derivation can be performed for Eq. (6), and hence the entire mean-field solution is exact in the limit of large system size.

IV. DISCUSSION

What does our solution of the Strauss model tell us? To begin with, it tells us the precise nature of and reason for the "degenerate state" observed by Strauss in simulation studies and by Burda *et al.* [19] in their perturbative calculations. Strauss's observations were correct—something special does happen to the model in the degenerate region. In fact, there is a first-order phase transition driven by the "field" parameter coupled to the number of edges in the graph. This also explains the breakdown of the perturbation expansion at this point, since such expansions typically break down at first-order transitions because of the corresponding pole in the free energy. The degenerate phase of the model is a high-density phase in which there is a large number of triangles in the graph, forming what appears to be almost a complete graph: the connectance of the network is close to 1 in this regime (Fig. 3).

More importantly, the first-order nature of the transition means there is a discontinuous jump in the density of triangles as we enter the degenerate state and thus there is no intermediate set of parameter values that will give the graph a moderate density of triangles as seen in real-world networks. While Strauss's model seems the most natural form for an exponential random graph model of transitivity, our results imply that it will in fact never be a good model of real-world networks with moderate clustering. One can of course reduce the value of the parameter α until we pass through the critical point so that the first-order transition disappears, in which case we recover smooth variation of the density of the graph with θ, but then the graph no longer has any significant clustering because of the small value of α.

These observations do not necessarily imply that exponential random graphs are incapable of mimicking networks with clustering; indeed they may present our best current hope for making clustered network models. Our results imply however that Strauss's original model with a single term in the Hamiltonian to encourage triangles must, at the very least, be augmented in some way in order to achieve this aim.

V. CONCLUSIONS

In this paper we have given a mean-field solution of Strauss's model of a network with clustering. Because of the intrinsically high-dimensional nature of networks, we believe this solution to be exact in the limit of large system size, which is the main case one is normally interested in. We have also performed Monte Carlo simulations of the model that confirm our solution to high accuracy. Our solution indicates that the model has no regime in which it displays moderate levels of clustering similar to those seen in real-world networks; presumably it will be necessary to introduce further terms into the Hamiltonian to avoid this pathology.

We believe exponential random graphs offer one of the most flexible tools for the modeling of general networks, and look forward to further developments. We hope that the formalism introduced here will serve as a practical starting point for a variety of problems.

Acknowledgments

This work was funded in part by the National Science Foundation under grant number DMS–0405348.

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] S. N. Dorogovtsev and J. F. F. Mendes, Advances in Physics **51**, 1079 (2002).
[3] M. E. J. Newman, SIAM Review **45**, 167 (2003).
[4] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Computer Communications Review **29**, 251 (1999).
[5] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[6] J. M. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, in *Proceedings of the International Conference on Combinatorics and Computing* (Springer, Berlin, 1999), no. 1627 in Lecture Notes in Computer Science, pp. 1–18.
[7] R. J. Williams and N. D. Martinez, Nature **404**, 180 (2000).
[8] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, Proc. Natl. Acad. Sci. USA **98**, 4569 (2001).
[9] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
[10] B. Bollobás, *Random Graphs* (Academic Press, New York, 2001), 2nd ed.
[11] P. Erdős and A. Rényi, Publications of the Mathematical Institute of the Hungarian Academy of Sciences **5**, 17 (1960).
[12] J. Park and M. E. J. Newman, Phys. Rev. E **70**, 066117 (2004).
[13] J. Park and M. E. J. Newman, Phys. Rev. E **70**, 066146 (2004).
[14] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Phys. Rev. E **69**, 046117 (2004).
[15] D. Strauss, SIAM Review **28**, 513 (1986).
[16] P. Holme and B. J. Kim, Phys. Rev. E **65**, 026107 (2002).
[17] K. Klemm and V. M. Eguiluz, Phys. Rev. E **65**, 036123 (2002).
[18] M. E. J. Newman, Phys. Rev. E **68**, 026121 (2003).
[19] Z. Burda, J. Jurkiewicz, and A. Krzywicki, Phys. Rev. E **69**, 026106 (2004).
[20] Z. Burda, J. Jurkiewicz, and A. Krzywicki, Phys. Rev. E **70**, 026106 (2004).
[21] G. Parisi, *Statistical Field Theory* (Addison-Wesley, Reading, MA, 1988).
[22] J. R. Banavar, M. Cieplak, and A. Maritan, Phys. Rev. Lett. **67**, 1807 (1991).
[23] Our calculations differ from those of Burda *et al.* in that Burda *et al.* study sparse graphs with finite mean degree where we study graphs with finite densities of edges. However, the solution we give should work in the sparse regime also, so direct comparison of the two calculations is valid.